Research project report: Modelling MAPK/ERK pathway using Ordinary Differential Equations

Przemysław Pilipczuk

A mechanistic model of MAPK/ERK signalling pathway was constructed using Ordniary Differential Equations and implemented into a pipeline for simulating and fitting parameters against experimental data. Three experiments were compared, and despite solid single-experiment fits, the cross-validation shows significant problems with generalization of fitted parameters across experiments.

October 13, 2025

Contents

1. Introduction

The mitogen-activated protein kinase/extracellular signal-regulated kinase (MAPK/ERK) pathway and its dynamics play a central role in determining cell fate in response to extracellular inputs. Different cell fates are linked to dynamics of the final node in the cascade, `ERK` kinases in particular. To understand how different dynamc patterns of their behavior arise, a mechanistical model of this pathway was constructed using ordinary differential equations (ODEs). Experimental data was obtained from a series of experiments including fibroblast cells transfected with optogenetic recetor tyrosine kinases (optoRTKs) constructs and an ERK-KTR reporter, that over the course of experiment were stimulated with different light patterns. Model parameters were estimated from this data, and cross-validation techniques were employed to evaluate generalizability of fitted parameters. Although individual experiments were fitted accurately, the resulting model failed to generalize across conditions, indicating need for further refinement.

2. Methods & Materials

2.1. Experimental Data

Data used for fitting the model was obtained from an experimental setup consisting of fibroblast cells transfected with an optoEGFR construct (optogenetic actuator) and being stimulated with various patterns of light. Data from 3 experiments were used in our pipeline, representing distinct light activation patterns. All experiments start with a 10 minute period of no actication, that we use for calibrating the baseline levels of activity.
First pattern that was investigated was a transient actuvation pulse for a given amount of time. The experiment was repeated with different durations, namely $t = \{0, 50, 100, 200, 500, 1000\}$ ms. This experiment was used during development for testing and verifying due to its simplicity.

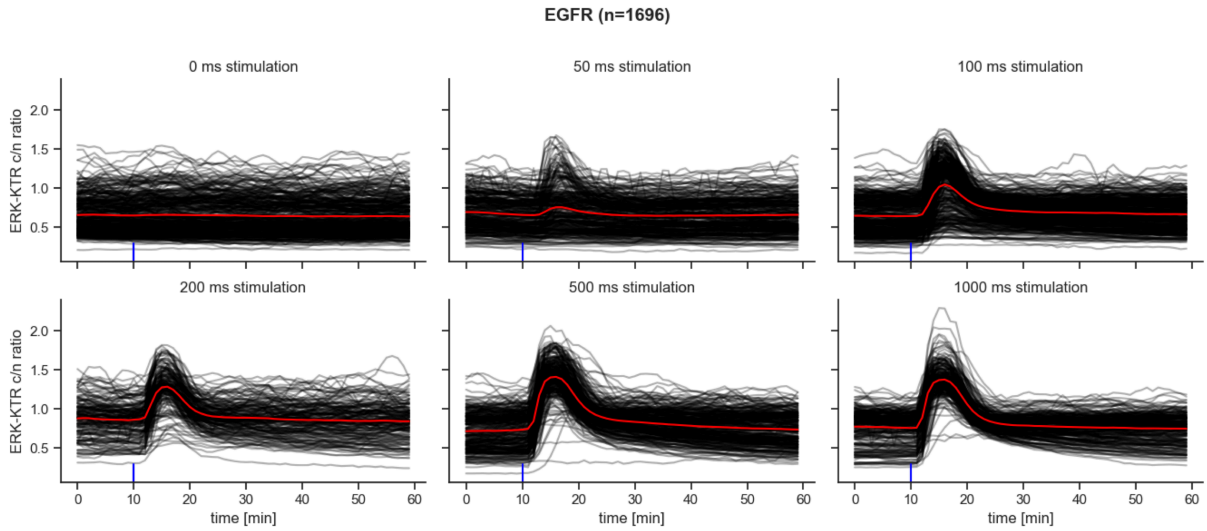

Figure 1: Single cell ERK-KTR activity in transient activation experiment. Each black curve represents a trajectory of a single cell. Red line is a median of all trajectories.

Second pattern that was incorporated was sustained activation. In this experiment, stimulation was started and sustained for a 140 minutes with three different power settings.

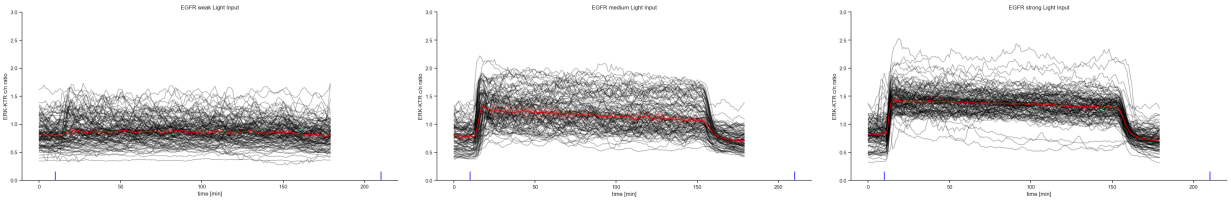

Figure 2: Single cell ERK-KTR activity in sustained activation experiment.
Each black curve is a single cell trajectory of ERK-KTR, red line describes median.

Third pattern incorporated into the pipeline was a ramp pattern. This pattern starts by running a light pulse every minute, and increasing the pulse widht with every subsequent actication, starting from 0 and ending at 700ms pulse width, resulting in overall ramp shape of the activation curve over time.

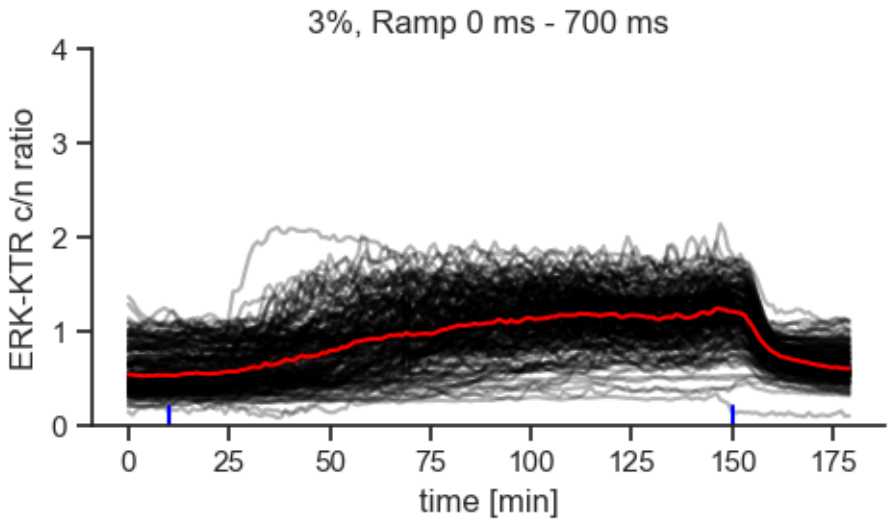

Figure 3: Single cell ERK-KTR activity in sustained activation experiment.
Each black curve is a single cell trajectory of ERK-KTR, red line describes median.

The development and initial testing of modeling pipeline was done using only transient activation experiment. Sustained and Ramp were incorporated as a means of providing cross-validation for the fitting process.

2.2. Model Structure

Finding the balance between simplicity and mechanistic accuracy is the key problem in model definition. Too simple model results in non-relevant findings, since it fails to capture complexity of the problem. Too complex model is harder to operate, and is prone to fragile to comlexity explosions, parameter unidentifiability, and poor choice of starting condition leading to nonsensical results.

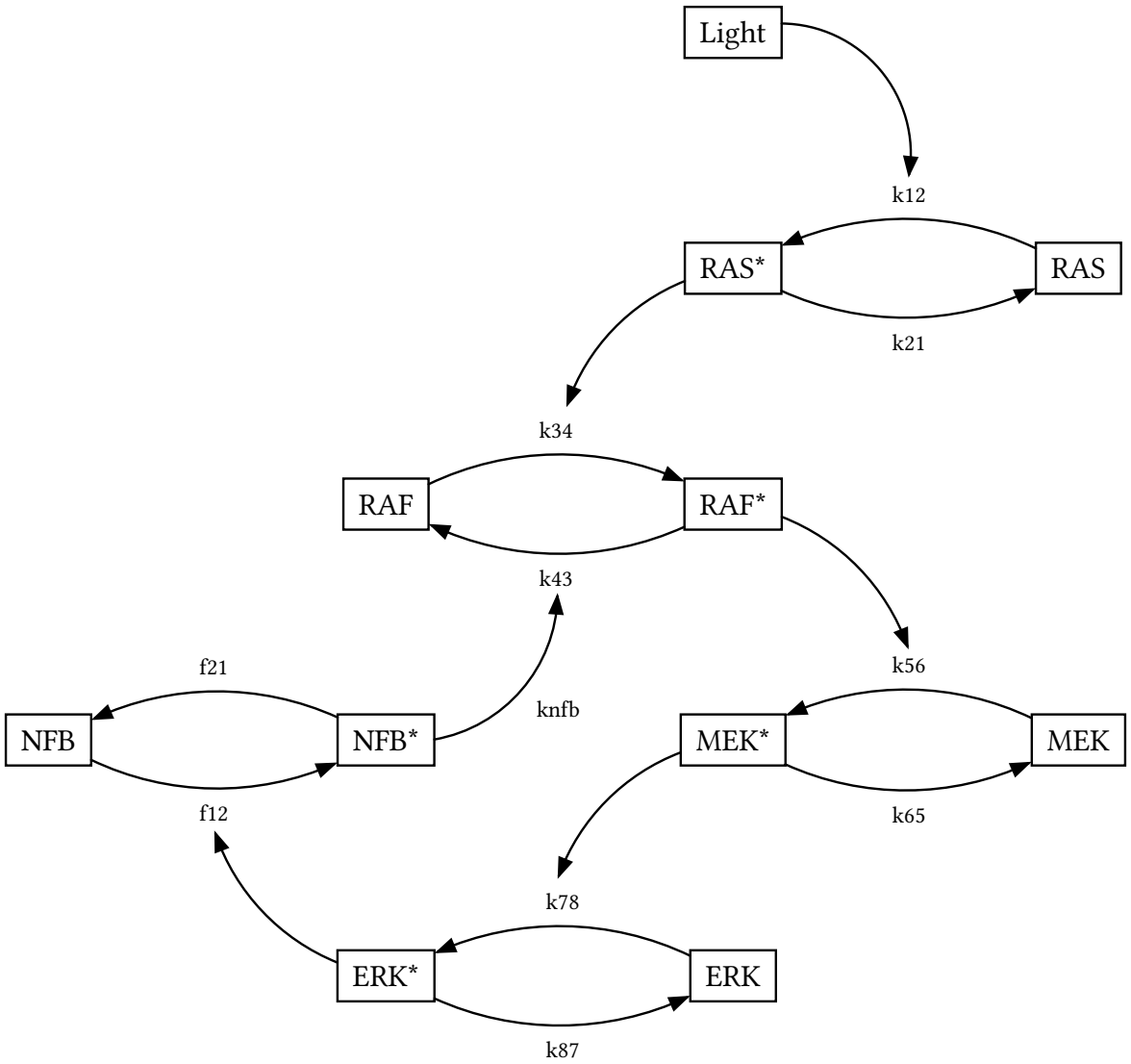

Figure 4: Diagram of the used model, representing simplified MAPK/ERK cascade.

Our chosen model starts at the level of RAS, and does not include the receptor layer. It treats light as a positive term in converting RAS to RAS*. The standard MAPK/ERK cascade of RAS-RAF-MEK-ERK is preserved We group whole families of kinases together and represent them as one vertex to preserve ontological grouping while minimising complexity of the model. Model also contains a negative feelback loop represented as a seperate NFB vertex in above graph. Modeling negative feedback as a separate state variable is not mechanistically accurate, but allows for a number of useful properties. It helps with model understandability, as negative feedback has clear and separate parameters that can be tweaked in isolation, while its existance as a state variable in the simulated model helps to track the contribution of negative feedback on the rest of the system. Another remarkable simplification of experimental setup comes from removal of reporter later (KTR-ERK complex). Presented model assumes that gathered experimental data coming from the reporter is a perfect proxy of actual state of ERK concentration in the cell.

2.3. Ordinary Differential Equations

Ordinary differential equations is a mathemathical framework for describing the change of one variable (dependant variable, here Y) over another variable (independent variable, here

X). Its simplest formulation is an expression that equates some mathematical expression to a derivative of our variable in question Y over the independent variable X.

$$\frac{dY}{dX} = ... \tag{1}$$

A solution to a differential equation is usually understood as obtaining a $Y(X)$ form, also called an *general solution*. General solutions can be obtained using an analytical solving process, the difficulty of which is heavily dependant on the specifc problem being solved. An often encountered problem with analytical solution approach is a problem that contains complex, nonlinearly coupled equations, which do not yield easily to this method. An alternative approach to solving a system that has these characteristics is a numerical one. This method relies on a provided initial conditions and a `X-step` resolution to simulate a single trajectory within an ODE system by evaluating the equations sequentially at a consecutive `X-steps` away from the provided starting point. This method, while providing a weaker form of solution, can deal with harder problems, including those that describe complex, nonlinear systems.
A load-bearing assumptions when picking up an ODEs for modeling dynamics of intracellucal concentrantions is that within the cell, the system is perfectly mixed (no spatial gradients of concentration occur). This assumption lets us avoid the complexities stemming from tracking chemical gradients within the cell based on position within it, which would be impossible to model using just an ODE system, as multiple additional independent variables arise from having to incorporate positional information.

Basic formulation of our model as a set of differential equations comes down to expressing each family of kinases as a state variable and expressing interactions between these kinases as a positive or negative terms in their respective differential equations.
Variables k and f are parameters of our model: those beginning with k generally concern the cascade, while those beginning with f correspond to the feedback mechanism, with a notable outlier of knfb, which is connected to both.

$$\begin{aligned}
\frac{d\mathrm{RAS}^*}{dt} &= \mathrm{light} * \left(\frac{\mathrm{RAS}}{K_{12} + \mathrm{RAS}}\right) - k_{21} * \left(\frac{\mathrm{RAS}^*}{K_{21} + \mathrm{RAS}^*}\right) \\
\frac{d\mathrm{RAF}^*}{dt} &= k_{34} * \mathrm{RAS}^* * \left(\frac{\mathrm{RAF}}{K_{34} + \mathrm{RAF}}\right) - (\mathrm{knfb} * \mathrm{NFB}^* + k_{43}) * \left(\frac{\mathrm{RAF}^*}{K_{43} + \mathrm{RAF}^*}\right) \\
\frac{d\mathrm{MEK}^*}{dt} &= k_{56} * \mathrm{RAF}^* * \left(\frac{\mathrm{MEK}}{K_{56} + \mathrm{MEK}}\right) - k_{65} * \left(\frac{\mathrm{MEK}^*}{K_{65} + \mathrm{MEK}^*}\right) \\
\frac{d\mathrm{NFB}^*}{dt} &= f_{12} * \mathrm{ERK}^* * \left(\frac{\mathrm{NFB}}{F_{12} + \mathrm{NFB}}\right) - f_{21} * \left(\frac{\mathrm{NFB}^*}{F_{21} + \mathrm{NFB}^*}\right) \\
\frac{d\mathrm{ERK}^*}{dt} &= k_{78} * \mathrm{MEK}^* * \left(\frac{\mathrm{ERK}}{K_{78} + \mathrm{ERK}}\right) - k_{87} * \left(\frac{\mathrm{ERK}^*}{K_{87} + \mathrm{ERK}^*}\right)
\end{aligned} \tag{2}$$

Figure 5: System of ODEs governing the simplified MAPK/ERK cascade.

The final assumption held was that on the timescale of observed experiments, the total amount of active and inactive form of a given molecule is constant (a “conserved moities” assumption). Such assumption allows for description of entire model using only half of the state variables,

by introducing each sum of forms as a constant in our model, thereby allowing us to refer the concentrations of active and inactive parts using a single concentration and a total (for example, instead of using RAS* and RAS, we can use RAS* and $\mathrm{RAS}_{\mathrm{total}} - \mathrm{RAS}^*$). This operation makes the simulation less computationally complex.

2.4. Modeling & Simulation pipeline

Model definition was done in symbolic form using SymPy in python (TODO: reference them), as a list of symbolic differential equations. The same tool was also used to lower the symbolic expression down to a numerical representation (`lambdify` function). A constructed model consists of a set of differential equations and parameter values. An effort was made to make this pipeline model-and-experiment agnostic and have low friction for implementing new models. This effort led to an architecture where each new model needs only to fulfill a simple interface for defining its equations and parameters, and each new experiment needs to define its own pattern of input and a parsing function for the data it generated.
An interactive environment for simulating behavior and exploring perturbations to parameters was constructed using marimo notebooks.
Parameter estimation was done using tools from SciPy library, namely `minimize` and `solve_ivp` to optimize loss function and to solve a numerical system respectively. The general scheme involved minimizing residual sum of squares for the data given in an experiment across all the groups in a given experiment.

Codebase is structured mainly around a `Model` class,That implements model definition, parameter estimation and simulating a trajectory given a stimulation pattern, parameters, and the initial condition. Each experiment has its own light-stimulation function and normalization and filtration of data.

3. Results

3.1. Parameter estimation from a single experiment

The preliminary fits over single experiments captures each individual dynamic response quite well in the broad strokes. However, a noticable pattern across these peaks is that the model rarely captures quantitative metrix such a $c_{\max}$, cannot account for changes in baseline, and has trouble accurately fitting scenarios with wide range of experimental groups.

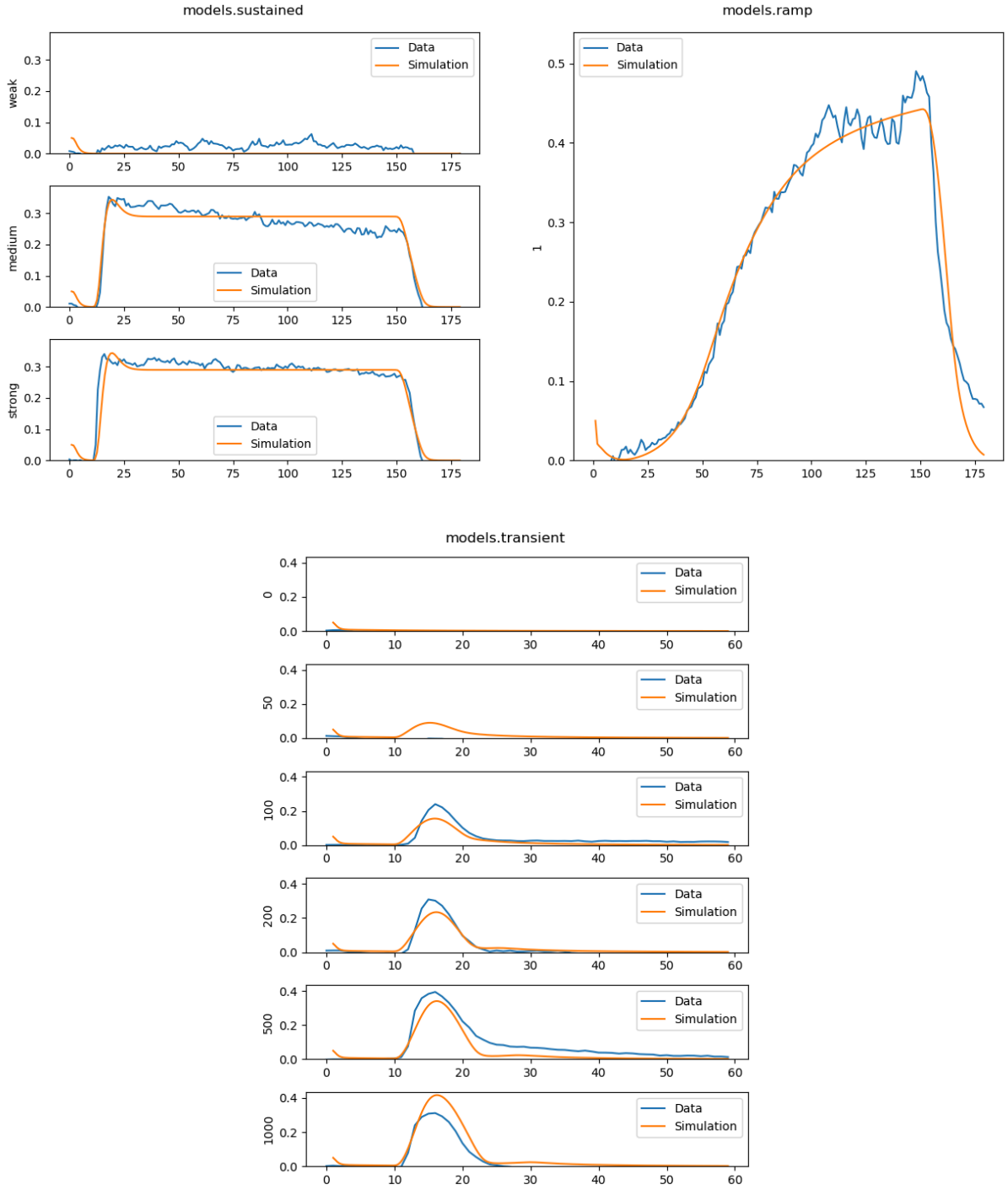

Figure 6: Results of 3 single-experiment fits.

3.2. Cross-validation

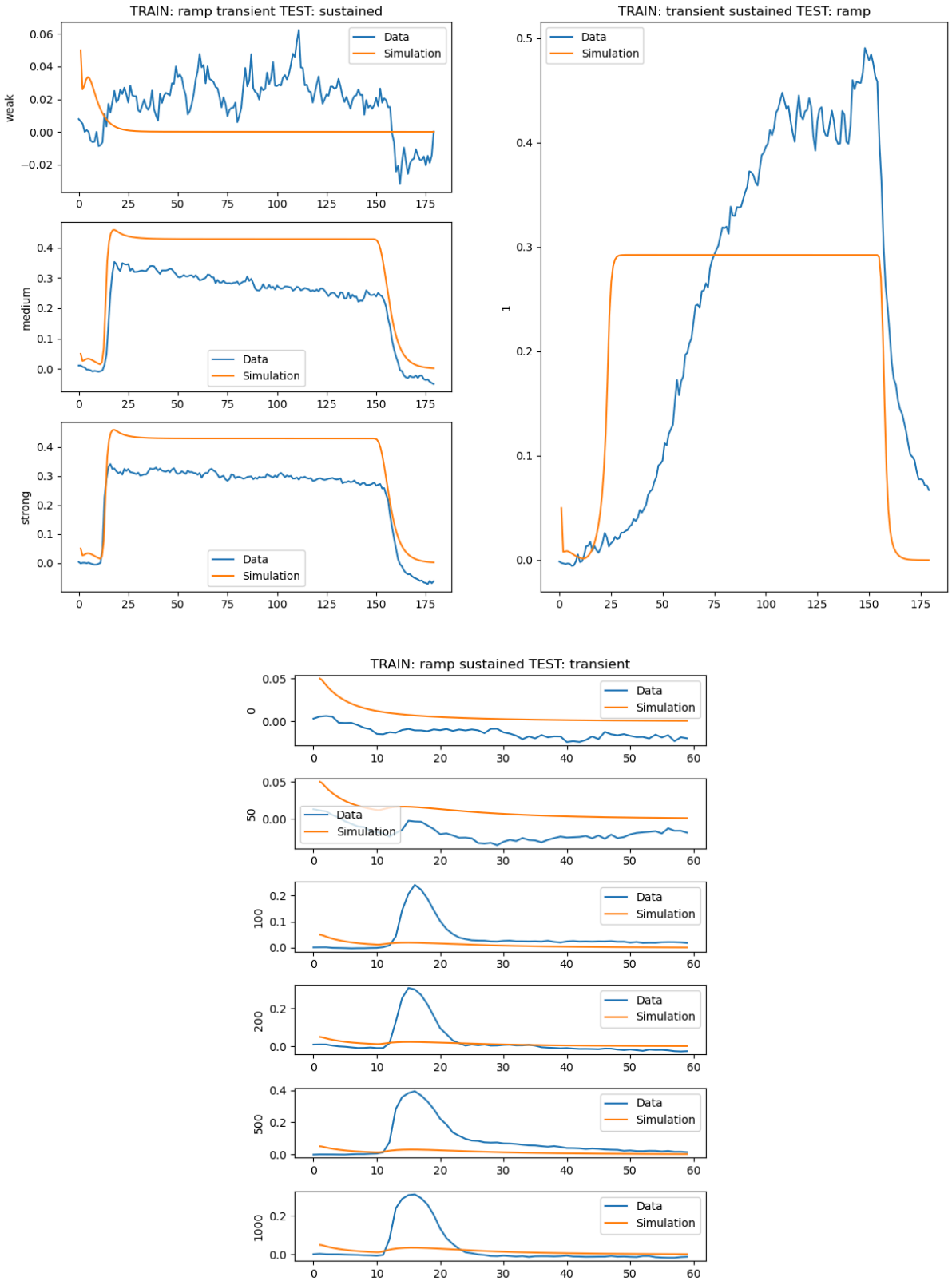

Figure 7: Results of 3-fold cross validation. There is observable lack of generalization across between experiments. Model trained on transient and ramp experiments and tested on sustained shows the best results in that it reproduces the shape of the activation we see in the data. However, the quantitative values outputted by the model are consistently overshooting the experimental observations. Other folds of cross-validation fail to reproduce the ERK activation dynamics at all.

4. Discussion

4.1. Time resolution trouble

During implementation of the parameter estimation regime, a technical difficulty was encountered. The experiments had been replicated multiple times with different parameter values for light stimulation, usually differing in stimulation time of a pulse of light, with values ranging from 50ms to 2000ms. At the same time, data for those experiments was collected with a rate of one sample per minute. Such difference in time ranges between sampling and input means that our experimental data will never be able to differentiate between the light stimulation duration groups within an experiment just by looking at the data (we would need to increase the sampling rate 1200 times to be able to discern signals that lasted 50ms, the most fine-grained light signal in our experimental data). To circumvent this problem, two solutions were contemplated.

1. An interpolation algorithm could have been used to synthetically create data points of the sampling rate we desire. This forces our new data points to inherit our assumptions about how the data is interpolated, which could introduce a systematic bias (as example, for linearity).
2. A proxy metric for a total light energy of a light pulse for a datapoint is introduced instead of a "binary" baseline. This allows for more flexibility but at a cost of having to create a realistically-scaling function that works for all kinds of experiments and across different levels of magnitude of light stimulation duration.

After theoretical evaluation of pros and const of both, a second solution strategy was picked, due to a better compatibility with the goal of having short training times.

5. Future work

Bibliography

Index of Figures